Homework 2

Megan Zhang

The purpose of this homework is to explore sampling distributions, to practice using the bootstrap to construct confidence intervals, and to gain more experience programming in R. Please fill in the appropriate code and write answers to all questions in the answer sections, then submit a compiled pdf with your answers through gradescope by 11:59pm on Sunday September 15th. **Note: you might find this homework is more challenging than the previous homework so please get started early.**

If you need help with any of the homework assignments, please attend the TA office hours which are listed on Canvas and/or ask questions on Piazza. Also, if you have completed the homework, please help others out by answering questions on Piazza.

Some tips for completing this homework are:

1) Make sure you conceptually understand the problems first before trying to write code for the solutions, i.e, if the problem is asking you to create a plot, draw a picture of the plot and think about the steps needed to get to the answer before writing down any code.
2) Several of the problems ask you to repeat previous problems with different parameter values. The best way to solve this is to do a careful job on the initial problem (e.g., label all the axes well, etc) and then copy your code over and adjust your parameters/answers (and sometimes it's possible to use a for loop over different parameter values to save on writing code).
3) Looking at the notes from class should be helpful

Some useful LaTeX symbols for the problem set are: μ, σ, $\bar{x}$, $\frac{a}{b}$

Problem 1: Exploring sampling distributions with simulations

As discussed in class:

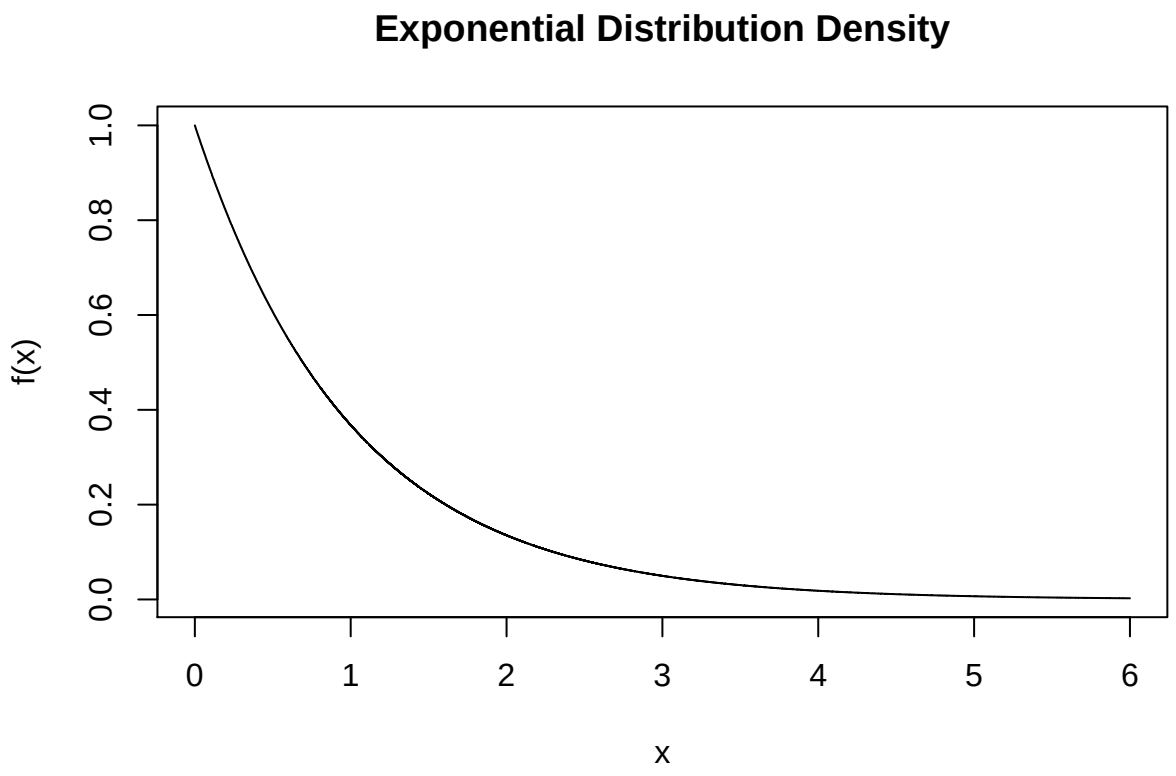
- A **statistic** is a number computed from a sample of data
- A **sampling distribution** is a probability distribution of a *statistic*; i.e., if we repeatedly drew samples of size n from some underlying distribution, and computed the same statistic on each sample, the distribution of these *statistics* is the *sampling distribution*.

The shape of the underlying distribution of data, and the shape of the sampling distribution for a statistic calculated from samples of data, are often quiet different. Below we explore this through simulations.

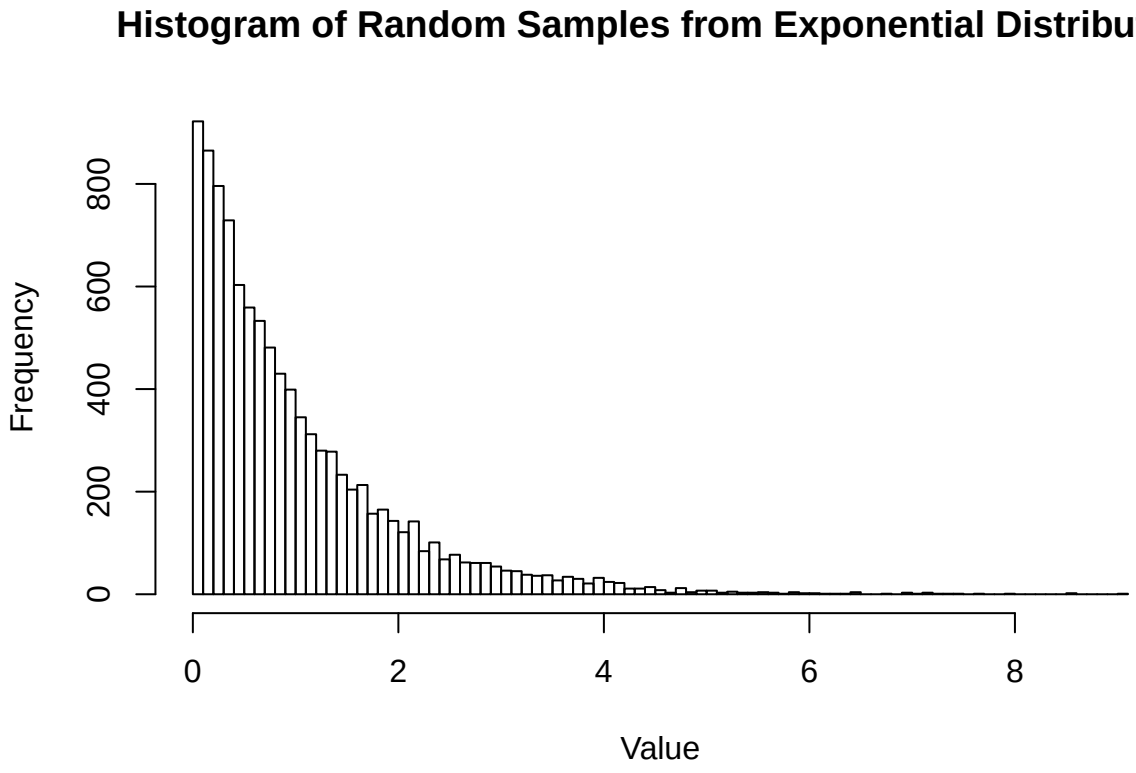
Problem 1.1: (15 points)

Let us examine data that comes from an exponential distribution with rate parameter $\lambda = 1$. Start by plotting the density for this exponential distribution using the dexp() function. Next randomly sample n = 10,000 points from the exponential distribution using the rexp(n) function and plot a histogram of these data (be sure to adjust the nclass argument to bin the histogram more finely). Also, calculate the mean, median and standard deviation of this randomly drawn data, and report the values of these statistics below using the LaTeX for the proper notation (use m for the notation for the median statistic). Finally, discuss whether the values of these statistics are what you would expect based on the values of the parameters of the exponential distribution (looking at the wikipedia entry to learn more about the parameters in the exponential distribution could be useful).

```
# plot the standard exponential density function
x <- seq(0, 6, by = .0001)  # x-values for plotting the exponential density function
y <- dexp(x)
plot(x,y, type = 'l', ylab = "f(x)", main = "Exponential Distribution Density")
```




```
# plot a sample of n = 10,000 points from this distribution
n <- 10000
hist(rexp(n), nclass = 100, xlab = "Value", ylab = "Frequency",
     main = "Histogram of Random Samples from Exponential Distribution")
```




```
# calculate some statistics from this sample
the_mean <- round(mean(rexp(n)), digits = 3)
the_median <- round(median(rexp(n)), digits = 3)
the_sd <- round(sd(rexp(n)), digits = 3)
```

Answers
The following are the statistics of central tendency:
Mean $\bar{x} = 1.003$
Median $m = 0.693$
Standard Deviation $s = 0.99$

The mean is about 1.00, median 0.693, standard deviation is also about 1.00. This matches the values based on the given parameters: Mean $=1\backslash\lambda = 1\backslash 1 = 1$, Median $= \ln(2)\backslash\lambda = \approx 0.693$, and standard deviation $= \sqrt{variance} = \sqrt{1\backslash\lambda^2} = 1$.

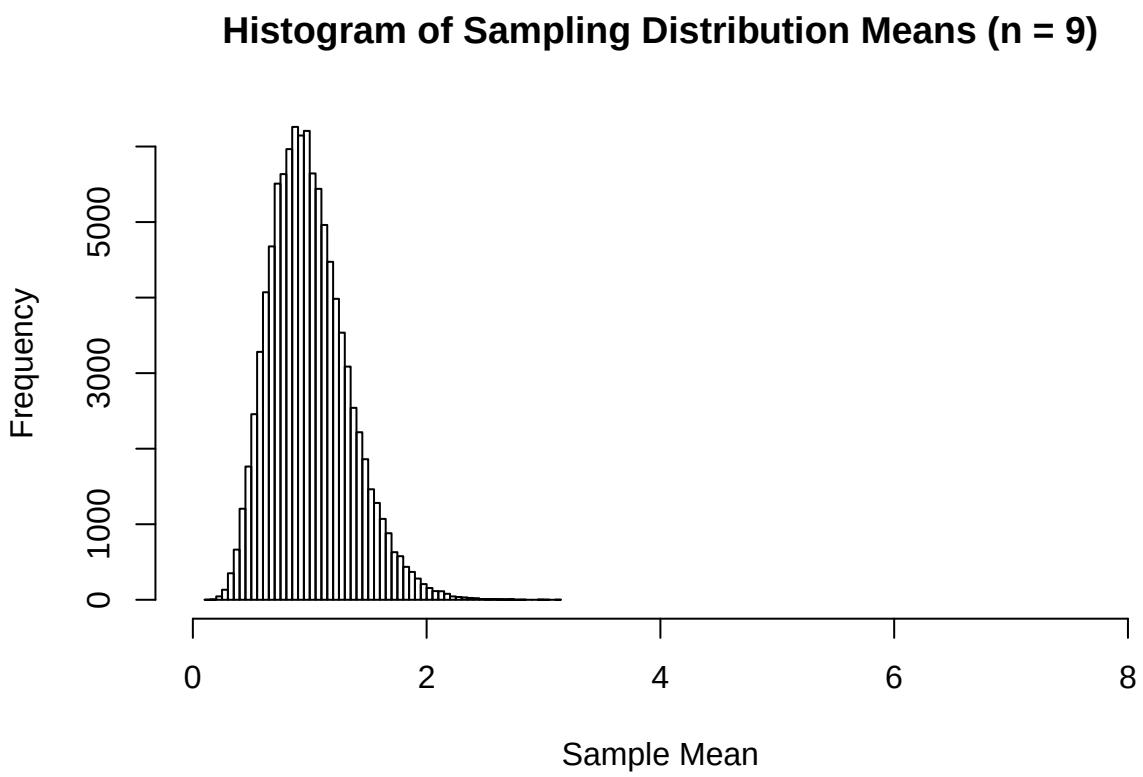
Problem 1.2: (15 points)

Now let's examine the *sampling distribution* for the mean statistic $\bar{x}$ when our underlying distribution is the exponential distribution. Use a for loop to create a sampling distribution that has 100,000 mean statistics, $\bar{x}$, using $n = 9$ points in each sample. Plot the distribution by creating a histogram of these sample statistic values, and set limits on the x-axis to be similar to those of the data distribution in problem 1.1 using the

xlim = c(lower_lim, upper_lim) argument. Finally, describe the shape of this distribution and report the standard error of this distribution.

```
sampling_dist <- NULL
for (i in 1:100000){
  sampling_dist[i] <- mean(sample(rexp(9)))
}

hist(sampling_dist, xlim = c(0,8), main = "Histogram of Sampling Distribution Means (n = 9)",
     xlab = "Sample Mean", ylab = "Frequency", nclass = 50)
```




```
the_SE <- round(sd(sampling_dist), 3)
the_SE
```

```
## [1] 0.334
```

Answers:

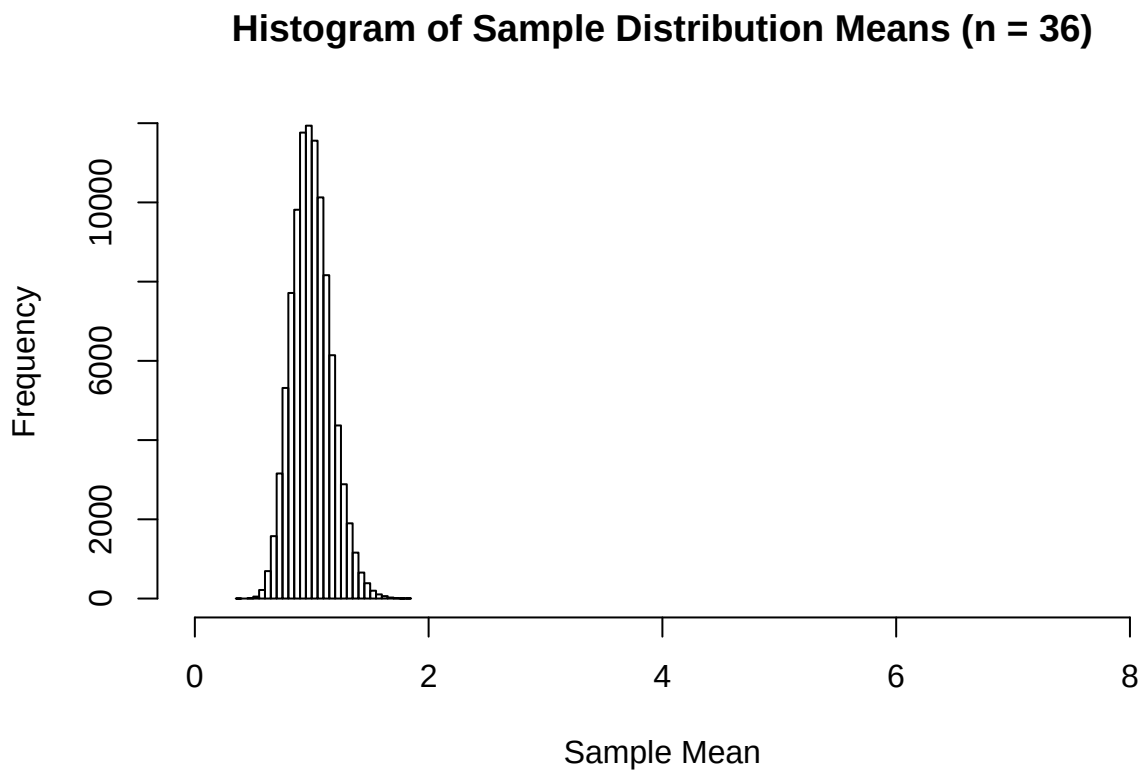
The distribution of the sample means is skewed right (the mean is larger than the median). The standard error is SE = 0.334.

Problem 1.3 (15 points)

Now repeat problem 1.2 using sample sizes of n = 36 and n = 144. Report the standard errors for $n = 9$, 36, and 144, and describe how the relationship between values for the standard error SE change with the different values of n. Also describe why it makes sense the SE would get smaller as n increases. Finally describe theoretical results (i.e., a formula) from intro stats that can account for the relationship between between the SE and n (hint: when you have an exponential distribution with rate parameter $\lambda = 1$, the standard deviation of this distribution is $\sigma = 1$).

```
sampling_dist <- NULL
for (i in 1:100000){
  sampling_dist[i] <- mean(sample(rexp(36)))
}

hist(sampling_dist, xlim = c(0,8), main = "Histogram of Sample Distribution Means (n = 36)",
     xlab = "Sample Mean", ylab = "Frequency", nclass = 30)
```



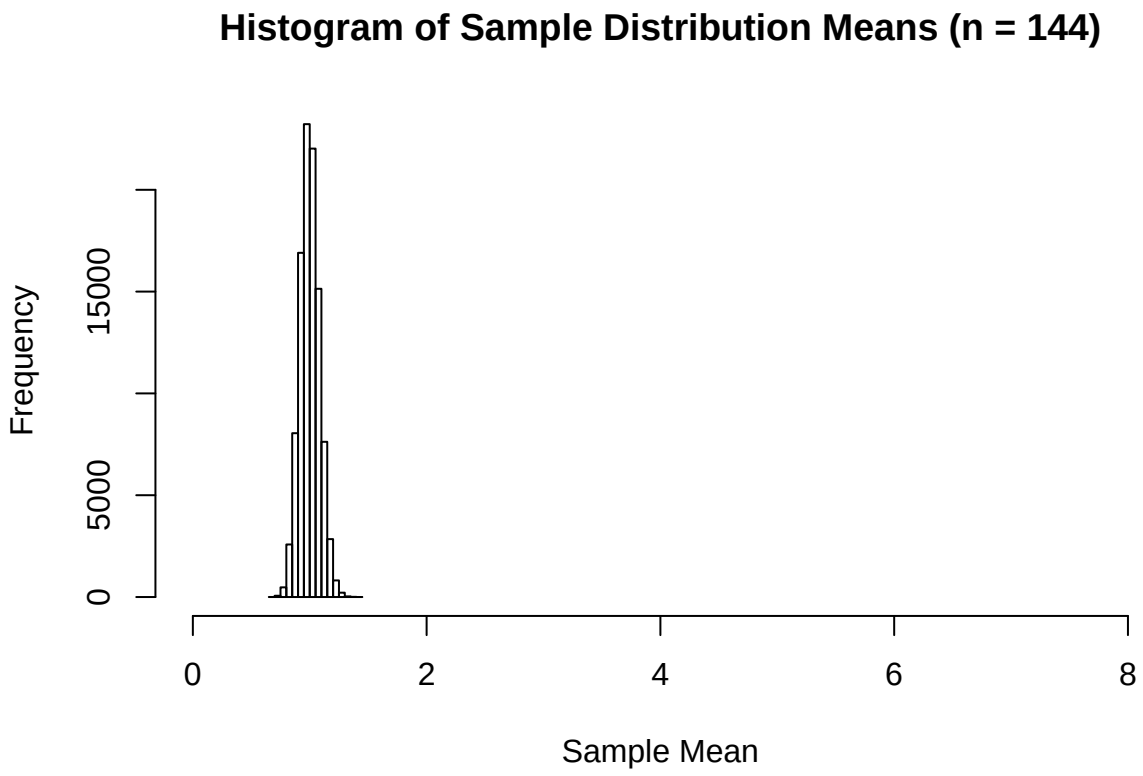

```
the_SE_1 <- round(sd(sampling_dist), 3)
the_SE_1
```

```
## [1] 0.167
```

```
sampling_dist <- NULL
for (i in 1:100000){
  sampling_dist[i] <- mean(sample(rexp(144)))
```

```
}

hist(sampling_dist, xlim = c(0,8), main = "Histogram of Sample Distribution Means (n = 144)",
     xlab = "Sample Mean", ylab = "Frequency", nclass = 20)
```




```
the_SE_2 <- round(sd(sampling_dist), 3)
the_SE_2
```

```
## [1] 0.083
```

Answer:

As n increases, the standard error decreases. This makes sense because according to the Law of Large numbers, as the sample size n increases the sample mean will converge to its expected value. This means the variance among sample means will decrease as they all become close to the true mean. We see this in the above histograms: as sample size increases, the means of the sample distribution become both more normally distributed and vary less. We also see this with our formula for standard error of the mean, which is $\sigma_M = \frac{\sigma}{\sqrt{N}}$, where n is the population size. As we increase the population of the sample, the standard error of the mean will decrease.

SE (n = 9) = 0.334.

SE (n = 36) = 0.167.

SE (n = 144) = 0.083.

Problem 2: Exploring bias correction in the formula for the variance statistic

In intro stats class you learned that the formula for the sample variance statistic is

$$s^2 = \frac{\Sigma_i^n (x_i - \bar{x})^2}{n-1}$$

One question that is often asked by students is why is the denominator in this formula n - 1 rather than just n. To examine this, let's create a sampling distribution of the variance statistic using a denominator of n - 1 and compare it to using a denominator of n.

Part 2.1 (10 points)

The function var() calculates the variance statistic from a data sample. Also, the function written below called var_n calculates the variance using a denominator of n rather than n-1. Create a sampling distribution using var() and var_n() when data comes from the standard normal distribution (using rnorm) for a sample size of n = 10. Plot histograms of these sampling distributions, and calculate the mean of these sampling distributions. Also use the abline(v = ...) function to plot a vertical line at the value of the parameter $\sigma^2 = 1$ (in red), and the value for the mean (expected value) of the sampling distribution (in blue). Then report below:

1) The shapes of these distributions
2) Whether the means of these sampling distribution equal the underlying variance parameter of $\sigma^2 = 1$.

Note: a statistic (i.e., estimator) is called *biased* if it's mean (expected value) does not equal the population parameter it is trying to estimate. Thus if the mean value of our sampling distribution does not equal the population parameter (in this case $\sigma^2 = 1$) then our statistic (estimator) is biased.

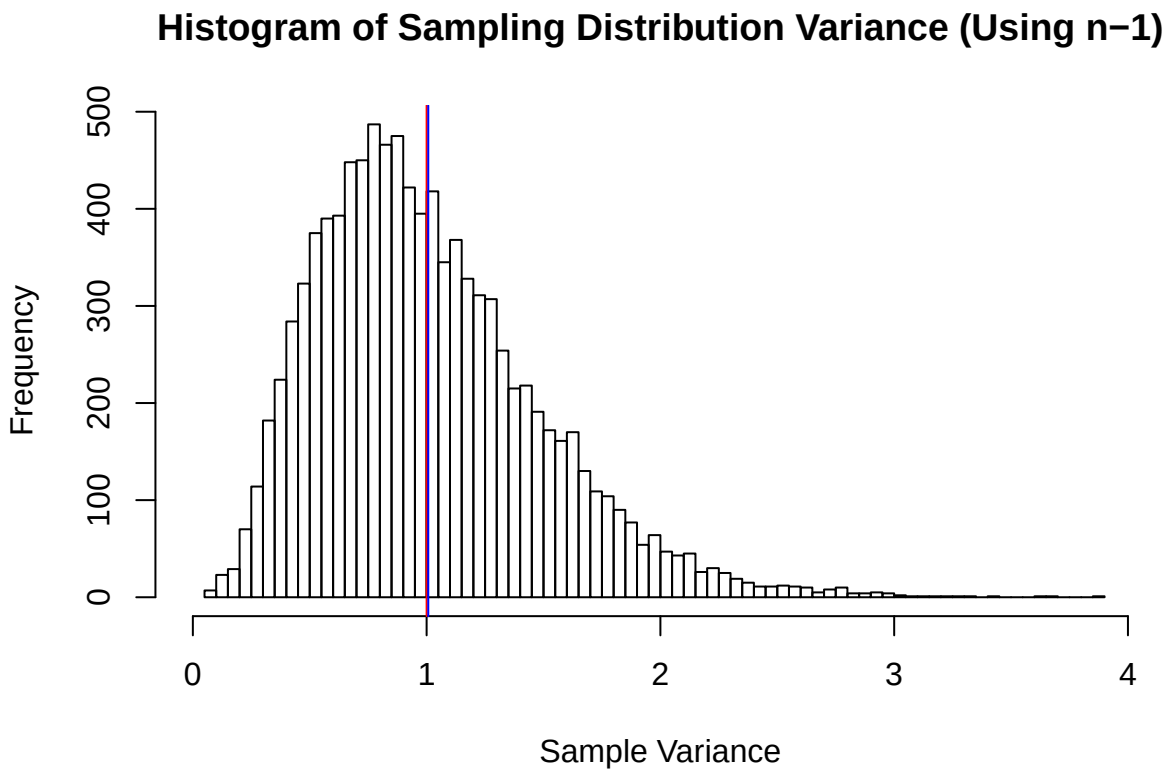
```
var_n <- function(data_sample){
  var(data_sample) * (length(data_sample) - 1)/length(data_sample)
}


# continue from here...
sampling_dist_1 <- NULL
for (i in 1:10000){
  sampling_dist_1[i] <- var(rnorm(10))
}

hist(sampling_dist_1, main = "Histogram of Sampling Distribution Variance (Using n-1)",
     xlab = "Sample Variance", ylab = "Frequency", nclass = 75, xlim = c(0,4))
mean_var <- mean(sampling_dist_1)
mean_var
```

```
## [1] 1.007032
```

```
abline(v = 1, col = "red")
abline(v = mean_var, col = "blue")
```

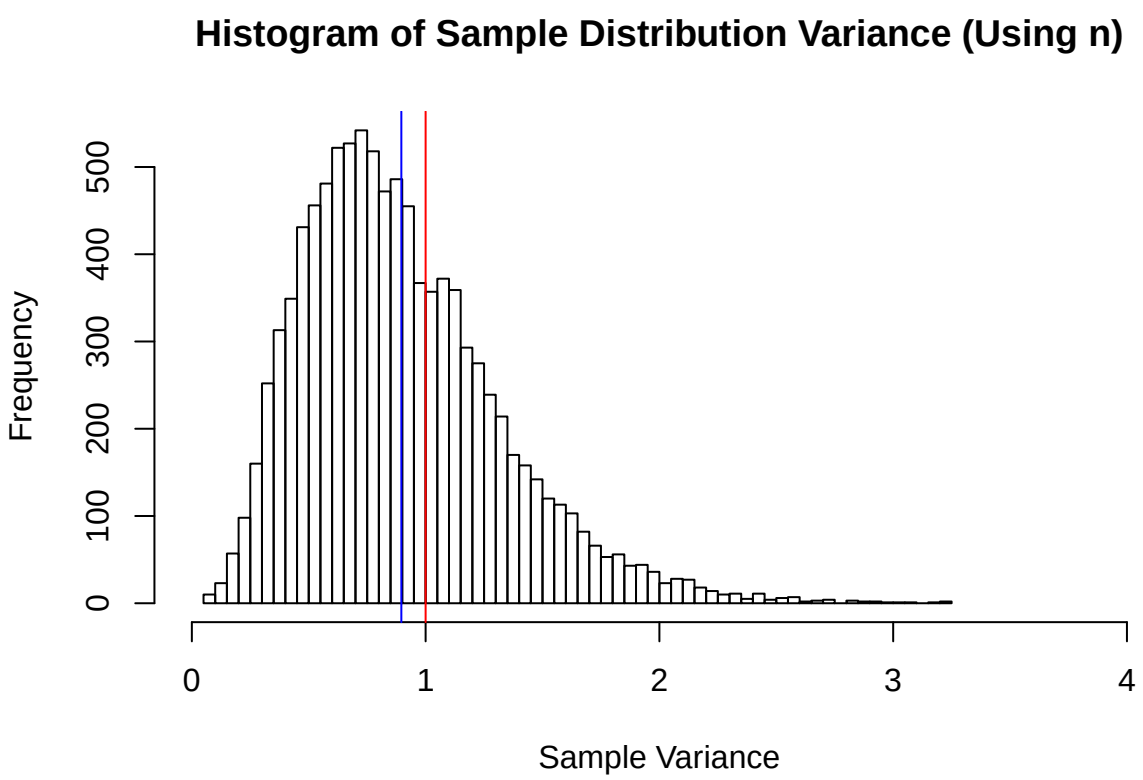



```
# continue from here...
sampling_dist_2 <- NULL
for (i in 1:10000){
  sampling_dist_2[i] <- var_n(rnorm(10))
}

hist(sampling_dist_2, main = "Histogram of Sample Distribution Variance (Using n)",
     xlab = "Sample Variance", ylab = "Frequency", nclass = 75, xlim = c(0,4))
mean_varn <- mean(sampling_dist_2)
mean_varn
```

```
## [1] 0.8961534
```

```
abline(v = 1, col = "red")
abline(v = mean_varn, col = "blue")
```



Answers:

Both are skewed right, but only mean of the distrubution using the function var() (n-1 is the denominator) is equivalent to the variance of 1. Here the blue and red lines are overlapping when we use n as the denominator to calculate variance. However, the variance using n as the demoninator is smaller than 1, so this sampling distribution is biased.

Part 2.2 (10 points)

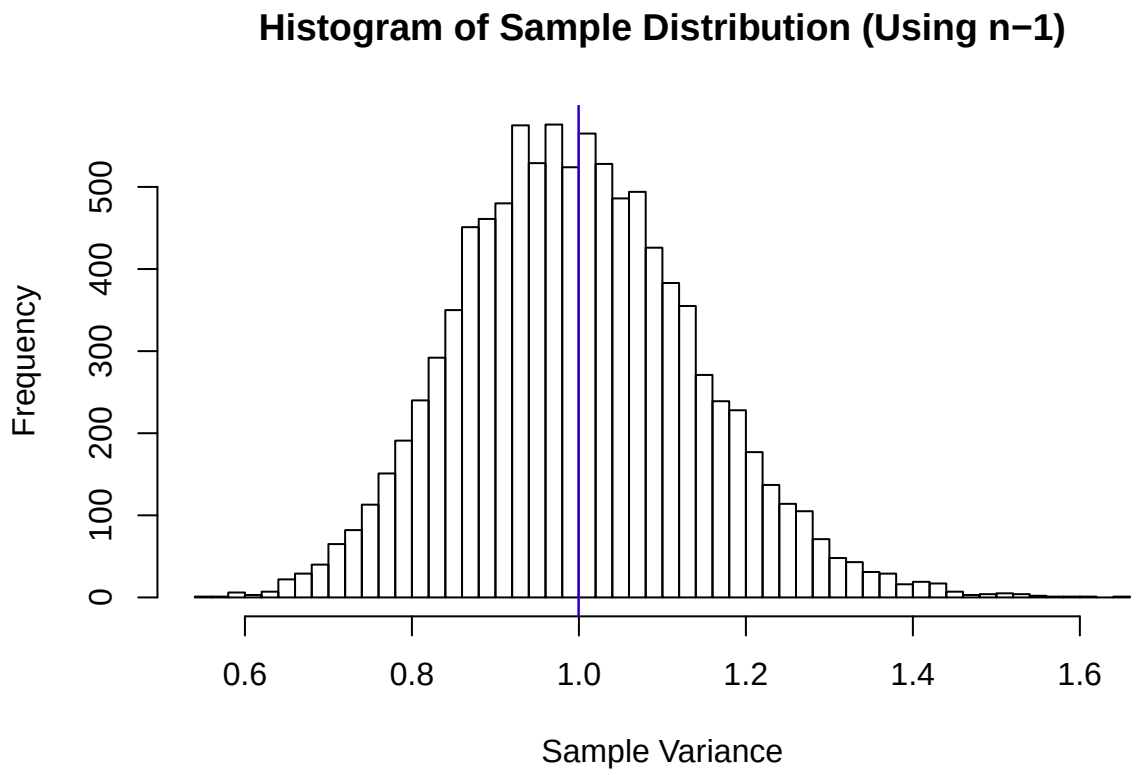
Repeat part 2.1 but using a sample size of n = 100. Do the answers to the questions posed in part 2.1 change?

```
sampling_dist_1 <- NULL
for (i in 1:10000){
  sampling_dist_1[i] <- var(rnorm(100))
}

hist(sampling_dist_1, main = "Histogram of Sample Distribution (Using n-1)",
     xlab = "Sample Variance", ylab = "Frequency", nclass = 75)
mean_var <- mean(sampling_dist_1)
mean_var
```

```
## [1] 0.9994331
```

```r
abline(v = 1, col = "red")
abline(v = mean_var, col = "blue")
```

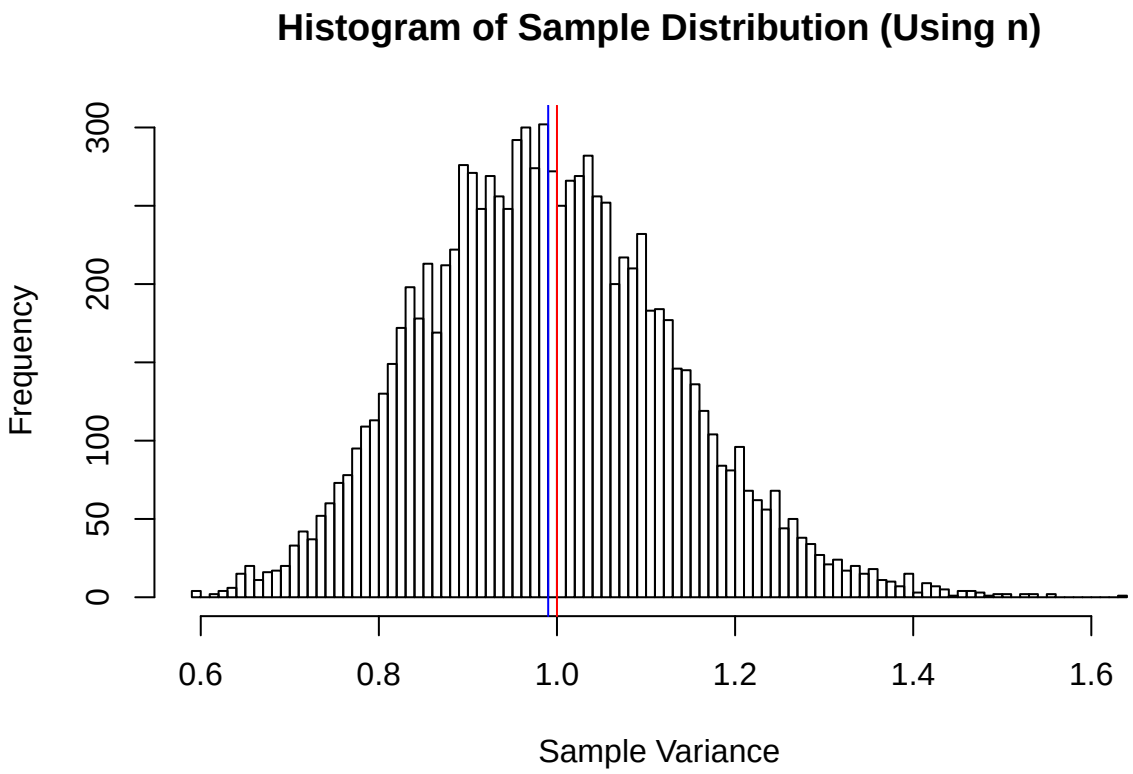



```r
# continue from here...
sampling_dist_2 <- NULL
for (i in 1:10000){
  sampling_dist_2[i] <- var_n(rnorm(100))
}

hist(sampling_dist_2, main = "Histogram of Sample Distribution (Using n)",
     xlab = "Sample Variance", ylab = "Frequency", nclass = 75)
mean_varn <- mean(sampling_dist_2)
mean_varn
```

```
## [1] 0.9901438
```

```r
abline(v = 1, col = "red")
abline(v = mean_varn, col = "blue")
```



Answers:

Adding to the sample causes the distribution to become more normally distributed. This is explained by the Central Limit Theorem. The sample variance calculated using n as the denominator is still less than the mean, but there is less difference due to the larger sample size.

Reflection (5 points)

Please reflect on how the homework went by going to Canvas, going to the Quizzes link, and clicking on Reflection on homework 2